Conformity Analysis of GTFS Routes and Bus Trajectories

Andreza Raquel M. Queiroz[1], **Veruska B. Santos**[1],
Dimas C. Nascimento[1,2], **Carlos Eduardo S. Pires**[1]

[1]Departamento de Sistemas e Computação
Universidade Federal de Campina Grande (UFCG)

[2]Unidade Acadêmica de Garanhuns
Universidade Federal Rural de Pernambuco (UFRPE)

{andrezaraquel,veruska}@copin.ufcg.edu.br

dimascnf@gmail.com, cesp@dsc.ufcg.edu.br

Abstract. *General Transit Feed Specification (GTFS) is a standard data format generated by transportation agencies of most of the cities worldwide to provide scheduled data of their services. Despite being the standard in the public transportation field, applications that consume GTFS data may face two problems: outdated versions, since some transportation agencies do not provide GTFS in the same frequency that transit changes; and discrepancy with positioning data sent by the buses. This paper provides a conformity analysis of GTFS routes and bus positioning data from multiple cities. We have found inconsistencies related to GPS route labels and GTFS routes. We also classify the conformity of bus trajectories and enumerate the main inconsistencies found in data analysis.*

1. Introduction

Several bus monitoring applications[1] use Global Positioning System (GPS) and General Transit Feed Specification (GTFS[2]) data as a source of supposedly accurate and official information about transit services. GTFS is a standard data composed of predefined bus routes represented by sequences of geographical points as well as other public transportation data. In turn, GPS bus data basically consists of records containing timestamp, vehicle number, bus route label, latitude and longitude. A set of GPS records with the same vehicle number ordered by timestamp is called a bus trajectory.

Bus monitoring applications usually face two crucial problems. The first one consists of outdated versions of GTFS data while the second problem is related to inconsistencies in the GPS data sent by buses. To address the first problem, the authors of [Wessel et al. 2017] developed a method to improve the accuracy of GTFS data using real-time GPS data sent by vehicles. Briefly, the method updates the GTFS data each time a significant change in a bus trajectory is detected. However, the authors do not deal with GPS inconsistencies which can introduce errors in the updated GTFS.

Ideally, the GPS route label should indicates the GTFS route that the bus is following. However, as mentioned before, there are inconsistencies in GPS data sent by

[1]Examples of bus monitoring applications: https://www.google.com/maps/, https://moovitapp.com/, https://www.ciomcg.com.br/

[2]https://developers.google.com/transit/gtfs/reference/

buses. These inconsistencies usually refer to a) missing GPS route label, probably due to device failures [Raymond and Imamichi 2016] and b) route label inconsistency, e.g., the route label indicated in the GPS data is different from the GTFS route that the bus is actually following. The problem of missing GPS route label was previously addressed in [Raymond and Imamichi 2016]. The authors have shown that the cosine similarity is an effective method to determine the routes followed by buses. In their experiments, using data from Rio de Janeiro, they compared the results generated using the cosine similarity with the route label sent by the buses. However, the authors neither considered the problem of outdated GTFS nor the second inconsistency in GPS route label, although their results were clearly affected by them.

In this work, we analyze the conformity of GTFS routes and bus trajectories. The analysis addresses both the outdated GTFS problem and GPS inconsistencies. For doing so, we use the cosine similarity method proposed in [Raymond and Imamichi 2016] to perform a more detailed study regarding the results generated by the method. Our main contributions are: a) we classify the conformity of bus trajectories with GTFS routes; and b) we enumerate the most frequent data inconsistencies found in our analysis. In addition, we demonstrate that, in some cases, the route determined by the cosine method is more likely to be the route that the bus is following than the route labeled in the GPS data.

2. Methodology

In order to analyze the conformity of GTFS routes and bus positioning data, we employ datasets of three Brazilian cities: City A (name omitted due to privacy requirements related to the usage of its bus GPS data), Curitiba and Rio de Janeiro. These datasets are summarized in Table 1. We downloaded Curitiba GTFS and GPS historical data from URBS (Urbanização de Curitiba S/A) web page. URBS is the agency that manages public transportation in Curitiba. Both datasets from Curitiba are publicly available[3]. City A data was made available by its public transportation agency. Rio de Janeiro GTFS and GPS data were provided by the authors of [Raymond and Imamichi 2016].

Table 1. Summary of datasets.

	GPS interval	# Bus trajectories	# GTFS routes
City A	Dec 03, 2018 to Dec 07, 2018	247	250
Curitiba	Aug 27, 2017 to Aug 31, 2017	697	235
Rio de Janeiro	Feb 15, 2016 to Feb 17, 2016.	5,376	375

To process the datasets and perform the conformity analysis, we followed the same steps of the approach proposed in [Raymond and Imamichi 2016] as follows.

2.1. Map matching of bus trajectories and GTFS routes

The first step is to apply map matching (MM) in bus trajectories and GTFS routes. MM is a process that integrates noise positioning data with the road network to obtain enhanced positions [Quddus et al. 2007]. In this work, we applied the GraphHopper MM algorithm that basically follows the approach described in [Newson and Krumm 2009]. GraphHopper implements the Hidden Markov Model (HMM) which is based on states, such that

[3]http://dadosabertos.c3sl.ufpr.br/curitibaurbs/, http://transporteservico.urbs.curitiba.pr.gov.br/index.php

the probability of the next state depends only on the current state. In the MM problem, the states are the road segments (i.e., a portion of a road between two intersections) of the road network. HMM map matching methods are known to be robust when dealing with GPS data which may contain measurement errors, as well as long and irregular intervals between measurements [Kubicka et al. 2018]. The code is publicly available[4]. The latest version of Open Street Map (OSM[5]) was used as the road network. Regarding Rio de Janeiro data, we used the MM output made available by [Raymond and Imamichi 2016].

2.2. Bag-of-roads transformation

The second step is to turn the MM outputs (i.e., the enhanced positions) generated in the previous step into vectors of roads. This process generates bag-of-roads (BoR) vectors to represent bus trajectories and GTFS routes. The BoR vectors maintain the same dimension in order to allow comparisons between them using classical similarity metrics, such as the cosine similarity. BoR vectors work analogously to the bag-of-words vectors for document classification. However, instead of counting words, each cell of a BoR vector represents a road segment and stores the frequency of a bus or route intersecting that segment.

We performed a modification in BoR vectors of bus trajectories in the sense that each cell simply indicates whether (or not) the bus intersected that segment. This modification brought better similarity results than the original BoR representation. This improvement occurs because bus trajectories are composed of various trips, all of them intersecting the same road segments. In turn, a trip is a sub-sequence of the trajectory that covers the route only once. In the original version proposed in [Raymond and Imamichi 2016], the authors assigned low similarity values to trajectories that are very similar to the route only because these trajectories are composed of several trips.

2.3. Route Identification

The last step is to identify the GTFS route of each bus. To this end, we compute the cosine similarity between BoR vectors representing bus trajectories and BoR vectors representing GTFS routes. The result of this step is a list of bus trajectories, each of them associated with its most similar route. Bus trajectories composed by GPS data whose route label does not correspond to an existing route in the GTFS were discarded. We consider more than one trajectory for the same bus if it follows more than one route and the label in its GPS data indicates this.

3. Conformity Analysis

We grouped the bus trajectories according to the similarity value between them and the corresponding GTFS route, determined using the cosine method. As a result, six groups were generated as depicted in Figure 1. Each group (represented by a box) indicates the conformity level of the bus trajectory. The horizontal axis refers to the similarity degree ranging from 0 to 1. The boxes above the horizontal axis include the cases in which the bus route label is in conformity with the found route. On the other hand, the boxes below the horizontal axis are related to cases where the bus route label is not in conformity with the found route.

[4]https://github.com/graphhopper/map-matching

[5]http://download.geofabrik.de/

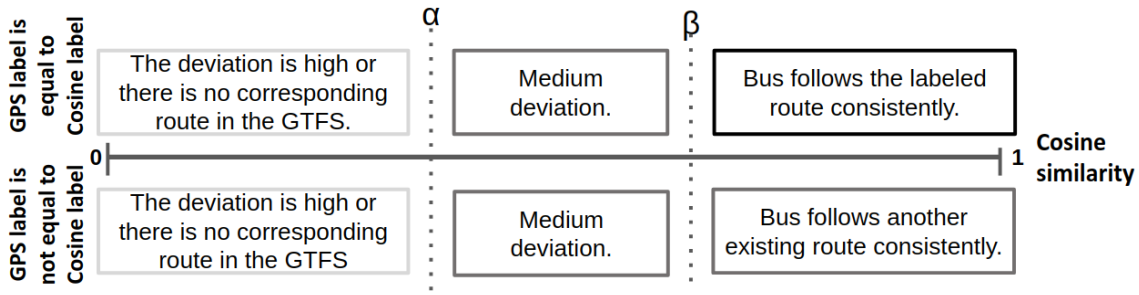

Figure 1. Conformity levels of bus trajectories based on their similarity with GTFS routes.

One level of conformity is assigned to each color of the boxes: a) black represents an ideal conformity level, i.e., the similarity of the trajectory with the GPS labeled route is high (similarity $\geq \beta$) and the GPS route label is equal to the found route; b) dark grey indicates that there is a certain discrepancy between bus trajectories and GTFS routes, i.e., the cosine similarity is not high ($\alpha <$ similarity $< \beta$) and/or the GPS route label is different from the found route; and c) light grey represents the worst cases of discrepancy with GTFS routes, i.e., the similarity between the bus trajectory and GTFS route is low (similarity $\leq \alpha$). α and β are configurable similarity thresholds used to classify the bus trajectories in levels. The threshold values should be chosen by a domain specialist. For the purpose of our analysis, we considered $\alpha = 0.4$ and $\beta = 0.7$, since these values partition the results consistently, generating cohesive groups.

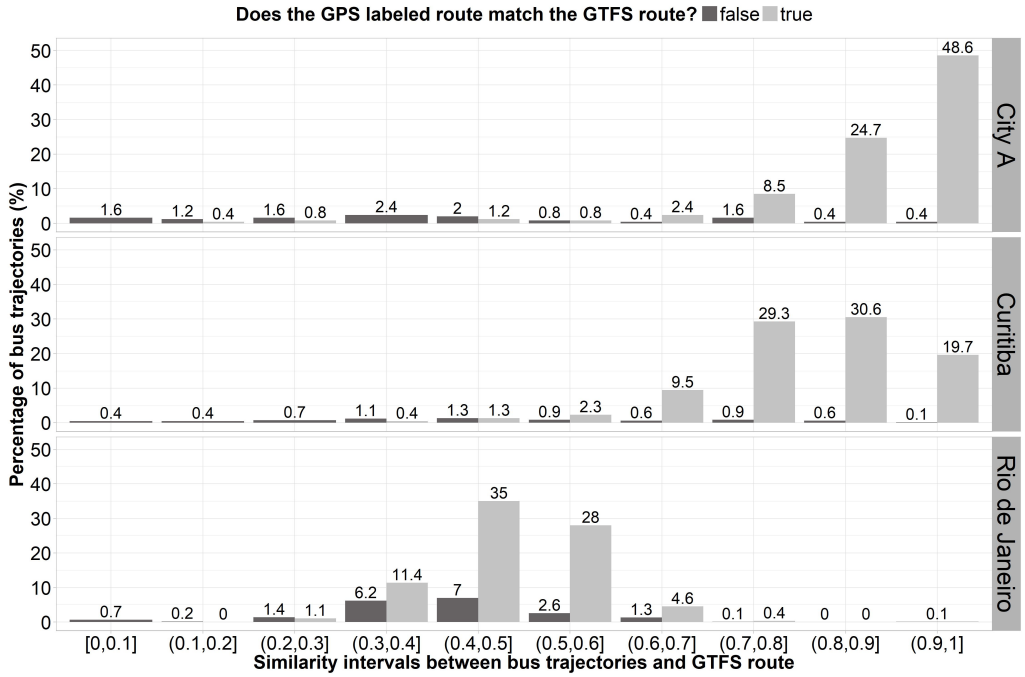

Figure 2. Percentage of bus trajectories that match (or not) the labeled GPS route, per similarity interval.

Figure 2 presents the percentage of bus trajectories that match (or not) the labeled GPS route, per similarity interval for each city. The horizontal axis represents the intervals of similarity whilst the vertical axis represents the percentage of bus trajectories.

The dark grey bars refer to the bus trajectories whose GPS route label is different from the identified route. In turn, the light grey bars refer to the bus trajectories whose GPS route label matches the identified route. Most buses from City A follow the labeled route consistently. However, 18.2% of its buses present a level of discrepancy with the GTFS. The same observation is verified for bus trajectories from Curitiba. On the other hand, Rio de Janeiro presents a different scenario. The bus trajectories are concentrated exactly in the intermediary region of similarity level, i.e., most of the trajectories show a medium deviation from the GTFS routes.

The low similarity levels associated with the bus trajectories can be explained by a variety of problems: GPS monitoring failure[6], outdated GTFS and/or bus route deviations. In the following, we enumerate the data inconsistencies generated by these problems and present examples of inconsistencies found in the analyzed datasets:

- **Inconsistency 1: the GPS route label is different from the route that the bus is following.** Figure 3 shows the trajectory of bus JC013. As it can be seen, the bus is actually following route 778 even though the GPS is labeled with route 776;
- **Inconsistency 2: the bus deviates partially from the route that it is actually following.** In Figure 3, bus LA002 never follows the highlighted portion of the route labeled on GPS;
- **Inconsistency 3: the bus follows more than one route, even though the GPS label indicates only one route.** In Figure 4, bus A29023 seems to be following both route 473 and route 441, but the GPS label indicates only route 473;
- **Inconsistency 4: the bus does not follow any of the GTFS routes.** In Figure 4, the trajectory of bus A48071 does not match any of the routes defined in the GTFS.

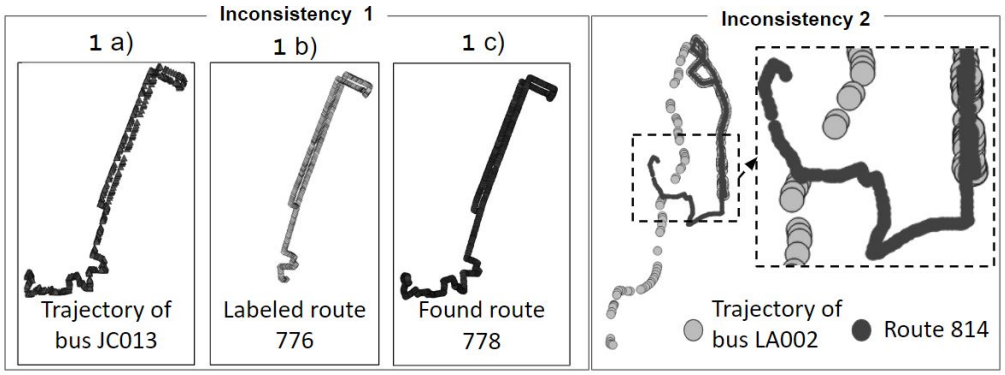

Figure 3. Examples of inconsistencies found in the Curitiba dataset. 1a: trajectory of bus JC013; 1b: route labeled in the GPS data; 1c: route found using the cosine method. The route determined by the cosine method is visually more similar to the bus trajectory than the GPS labeled route. Inconsistency 2: the bus never follows the highlighted part of the route.

4. Conclusion

GTFS is a standard that facilitates the exchange of public transportation data and should be considered the gold standard in the field. However, we demonstrated that buses oper-

[6]https://g1.globo.com/rj/rio-de-janeiro/noticia/fora-do-ponto-mais-da-metade-dos-onibus-do-rio-tem-falha-no-monitoramento-por-gps.ghtml

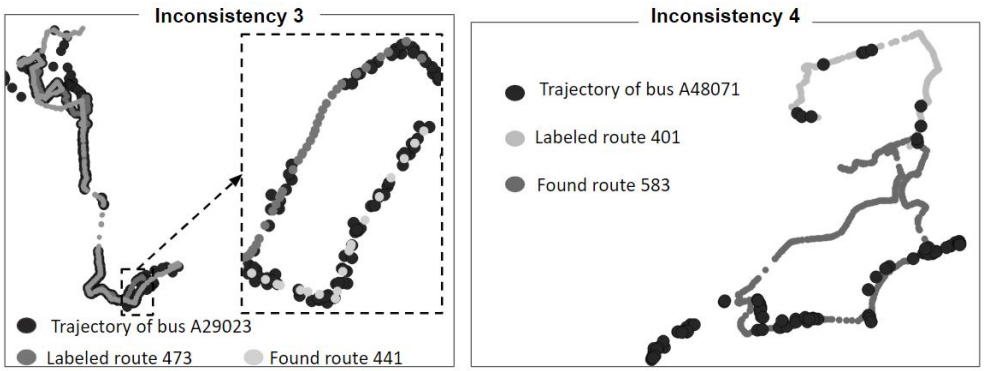

Figure 4. Examples of inconsistencies found in the Rio de Janeiro dataset. Inconsistency 3: the highlighted part indicates that bus A29023 is following both routes 473 and 441. Inconsistency 4: bus A48071 is neither following the found route (583) nor the labeled route (401).

ating in diverse cities do not always follow the programmed route, showing a certain inconsistency with the GTFS. Some inconsistencies are more serious and should be treated immediately, such as the incorrect GPS route label. Other inconsistencies, such as the bus following more than one route or deviating from its predefined route could be avoided with a more effective strategic plan in the GTFS creation and real-time supervision of bus GPS data. The categorization of conformity presented in this work can be extended by a specialist in public transportation, according to its interests. Future work will focus on generate GTFS routes based on bus trajectory data. Thus, we can ensure that GTFS data is updated and also it is in conformity with GPS data.

Acknowledgment

This research was partially funded by INES 2.0, FACEPE grant APQ-0399-1.03/17, CAPES grant 88887.136410/2017-00 and CNPq grant 465614/2014-0.

References

Kubicka, M., Cela, A., Mounier, H., and Niculescu, S.-I. (2018). Comparative study and application-oriented classification of vehicular map-matching methods. *IEEE Intelligent Transportation Systems Magazine*, 10(2):150–166.

Newson, P. and Krumm, J. (2009). Hidden markov map matching through noise and sparseness. In *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 336–343. ACM.

Quddus, M. A., Ochieng, W. Y., and Noland, R. B. (2007). Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation research part c: Emerging technologies*, 15(5):312–328.

Raymond, R. and Imamichi, T. (2016). Bus trajectory identification by map-matching. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 1618–1623. IEEE.

Wessel, N., Allen, J., and Farber, S. (2017). Constructing a routable retrospective transit timetable from a real-time vehicle location feed and gtfs. *Journal of Transport Geography*, 62:92–97.